TABLE VI. Comparison of Results for the *CAPLUS* and *SICHO* Models With Exact Secondary and Tertiary Constraints

PDB Name	Number of Residues	Type	Number of Constraints	cRMSD in Å from the *SICHO* Model[a,b]	cRMSD in Å from the *CAPLUS* Model[a]
1gb1	56	α/β	8	3.4	3.3
1ctf	68	α/β	10	3.2	4.2
1pcy	99	β	46	3.8	3.5
1pcy	99	β	25	4.9	5.4
1pcy	99	β	15	5.7	---
2trx	108	α/β	30	3.1	3.4
2trx	108	α/β	16	3.5	---
4fab	113	β	27	4.4	---
4fab	113	β	16	5.9	---
3fxn	138	α/β	35	4.1	3.9
3fxn	138	α/β	20	4.1	---
1mba	146	α	20	4.3	5.9
Atim	247	α/β	62	5.1	---
Atim	247	α/β	50	6.0	---
Atim	247	α/β	36	6.7	---

[a] Average cRMSD of the Cα over an isothermal stability run.

[b] The average cRMSD is reported from structures obtained after the SICHO model has been mapped into the CAPLUS model and relaxed.

Somewhat surprisingly, there is no significant difference between the average quality of the rebuilt Cα chains and that roughly estimated from a simple linear combination of three successive side chain centers of mass. This shows that the side chain model is consistent with the CAPLUS model used previously. The Cα reconstruction process employed here neglects all the long-range interactions (except of course the target harmonic constraints), and was is done for the sake of computational efficiency.

Comparison With Other Work

As mentioned above, there have been several other attempts to use known secondary structure and some tertiary constraints in the prediction of protein three-dimensional structures. However, the closest studies of other workers who used both known secondary structure and exact tertiary constraints are those of Smith-Brown and coworkers and Aszodi and Taylor. Smith-Brown *et al.* reported the examination of a number of proteins. By way of example, flavodoxin, a 138 residue α/β protein, was folded to a structure whose backbone cRMSD from native was 3.18 Å for 147 constraints. In contrast, with just 20 constraints, here structures whose cRMSD from native is 4.2 Å were generated. Similarly, for 3fab, 90 constraints were reportedly required to produce a model whose cRMSD was said to be 4.6 Å. For 4fab in the present approach, the use of just 27 constraints yielded a model whose cRMSD was 4.4 Å. The reported requirement for a large number of constraints was likely due to the lack of knowledge-based, protein-like background potential.

Another effort to predict the global fold of a protein from a limited number of distance constraints is due to Aszodi *et al.*[5] In general, they find that to assemble structures below 5 Å cRMSD, on average, typically more than N/4 constraints are required, where N is the number of residues. Even then, the method reported by Aszodi *et al.* had problems selecting out the correct fold from competing alternatives. While their best folds are of acceptable accuracy, the competing misfolded structures could be disregarded based on energetic considerations. In contrast, in the simulations presented here, the nativelike fold was easily detected as the lowest energy structure, and just N/7 constraints were required to produce structures of comparable accuracy.

The MONSSTER algorithm uses the CAPLUS model,[6] and also employs a reduced lattice model of protein, a background, knowledge-based force field, and a simulated thermal annealing Monte Carlo procedure for fold assembly. Using MONSSTER, about N/4 constraints are required to assemble β-type and α/β-

proteins, while helical proteins required N/7 constraints. Here, for a representative set of proteins, all types of folds can be assembled with knowledge of, on average, N/7 tertiary constraints. In addition, the results are less sensitive to the distribution of constraints. For example, in the case of the 18-55 fragments of 6pti in the CAPLUS model, the cRMSD was about 6-8 Å for the different sets of constraints. In contrast, in the side-chain-based model, for all sets of examined constraints, it is about 4 Å. Furthermore, as evidenced by the ability to fold the 247-residue TIM from a fully extended state, much larger systems can be treated. With an increasing number of long-range constraints, the accuracy of assembled structures increases and is consistently better than for previously reported methods. In addition, the resulting models are found to be less sensitive to the constraint distribution.

The instant invention also offers the advantage of speed. For small proteins, the algorithm is essentially interactive. It takes about 5-10 minutes of CPU time on a contemporary workstation to assemble the relatively complex motif of a 68-residue 1ctf fragment. Since the cost scales approximately as N^3, assembly of larger structures requires more time. Thus, a myoglobin folding simulation requires about 2 hours of CPU time.

CONCLUSION

This example demonstrates that the invention provides a powerful new model for the assembly of three-dimensional protein structures from known secondary structure and a small number of tertiary constraints. While the model only explicitly considers side chain centers of mass of the amino acid residues comprising the protein being studied, the effect of backbone atoms is implicitly built into the model force field, which also exploits the structural regularities seen in protein structures. Thus, the invention is fully compatible with more complex models that employ a larger number of united atoms per residue. In all respects, the invention compares favorably with previous approaches having a similar goal: the assembly of tertiary structure from loosely encoded secondary structural biases and a